NOISE REDUCTION IN DUAL-MICROPHONE MOBILE PHONES USING A BANK OF PRE-MEASURED TARGET-CANCELLATION FILTERS

Zbyněk Koldovský[1,2], Petr Tichavský[2], and David Botka[1]

[1]Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
[2]Institute of Information Theory and Automation,
P.O.Box 18, 182 08 Prague 8, Czech Republic

ABSTRACT

In this paper, a novel method of noise reduction for dual-microphone mobile phones is proposed. The method is based on a set (bank) of target-cancellation filters derived in a noise-free situation for different possible positions of the phone with respect to the speaker mouth. Next, a novel construction of the target-cancellation filter is proposed, which is suitable for the application. The set of the cancellation filters is used to accurately estimate the noise of the environment, which is then subtracted from the recorded signal via standard Wiener filter or a power level difference method. Experiments with recorded data show a good performance and low complexity of the system, making it possible for an integration into mobile communication devices.

Index Terms — Noise Reduction; Speech Enhancement; Dual-Channel; Target-Cancellation Filters; Wiener Filter

1. INTRODUCTION

Noise suppression from a voice of a mobile-phone user is a hot topic of audio signal processing since there are billions of users over the world. Until recently, mobiles have been equipped by one microphone, so single-channel methods [1, 3] have been applied. However, the immense progress already allows the integration of two or more microphones into one mobile. A special attention is therefore paid to dual-channel processing methods. Two microphones could be used for the noise suppression, which is the target application focused in this paper, but also for other entertainment or multimedia applications such as stereophonic audio recording.

Most methods enhance the speaker voice by suppressing all the other sounds (the noise) from the noisy voice recording, so any information about the noise is the key need. To this end, the diversity between channels can be exploited. Some methods estimate noise power spectral density by detecting noise-only or noise-dominant time-frequency intervals [4, 5, 6, 7]. The coherence function between signals from two microphones is used in [8] to design a noise reduction filter. Blind source separation based on ICA can be used to separate the voice and noise [9] and to exploit the separated signals in a post-processing stage [10].

Popular methods of noise suppression are adaptive beamformers having the structure of the Generalized Sidelobe Canceller (GSC) [11, 12, 13]. In these methods, a reference noise signal is obtained as an output of a block (called the Blocking Matrix) which is, in fact, a target-cancellation filter (CF) that cancels the speaker voice but passes the noise. Provided that the CF performs well, the noise can be observed even during intervals of the speaker activity, hence its subsequent suppression can be very efficient.

However, there are two major problems. First, the CF must be designed according to the position of the speaker, which is rarely fixed. Moreover, the propagation of sound in real environment (reflections and reverberations) should be taken into account. The second problem is that the spectrum of the passed noise is changed by the CF in an unknown way.

Pioneering beamformers [14] assume free-field conditions and design the CF based on an estimation of direction-of-arrival of the dominant source. More advanced methods [15, 16, 17] take real acoustic into account but require speaker-only measurements to compute the CF for the current speaker position. The spectrum of the CF output is usually corrected in an adaptive noise canceler by a least-mean-squares adaptive filter [22].

In this paper, we propose a novel noise reduction method suitable for mobile phones, where the position of speaker is mostly limited to the immediate vicinity of the microphones. The method uses a set (bank) of cancellation filters that were computed in advance under noise-free conditions for the most probable positions of the speaker. We also propose a novel cancellation filter design, which minimizes a distortion of the noise spectrum.

We compare the proposed method with the state-of-the-art method of Jeub et al. [4] presented last year at this conference. The latter method is based on Power Level Differences (PLD)
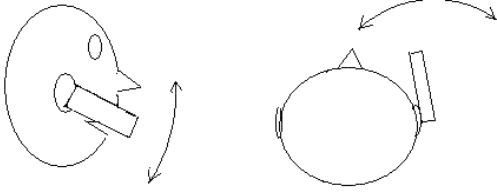
Fig. 1. Range of typical positions of the mobile phone for preparation of the cancellation filter bank.

and assumes that the secondary microphone is placed on the rear side of the mobile. Our method can be designed for any microphone arrangement. In comparison to PLD, it achieves better perceptual quality and is able to work in difficult scenarios where SNR is lower than 0 dB.

The paper is organized as follows. In Section 2, a construction of the CF bank is described. In Section 3, the noise suppression algorithm is proposed, which uses the filter bank. Section 4 presents experiments and Section 5 concludes the paper.

2. CANCELLATION FILTER BANK

Each filter in the bank is measured and computed for one particular position of the mobile with respect to the speaker. The positions should cover a range of expected positions of the mobile during an ordinary telephone conversation which is schematically shown in Fig. 1. For each position, an utterance of a speaker should be recorded in a quiet room. We rely on the empirical fact that the cancellation filters mostly depend on the construction of the mobile phone and its position w.r.t. speaker's head, but are less dependent on other objects.

2.1. Target-Cancellation Filters

A dual-channel recording of a target source during which its position is fixed is described by

$$x_{\mathrm{L}}(n) = \{h_{\mathrm{L}} * s\}(n) + y_{\mathrm{L}}(n),$$
$$x_{\mathrm{R}}(n) = \{h_{\mathrm{R}} * s\}(n) + y_{\mathrm{R}}(n) \quad (1)$$

where $n = 1, \ldots, N$ is the time index, $*$ denotes the convolution, $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are, respectively, the signals from the left and right microphone, $s(n)$ is the target signal, and $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ are noise signals (further referred to as "noise"). $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ denote the microphone-source impulse responses.

An ideal filter that cancels the target signal s, generally, consists of two non-zero SISO filters g_{L} and g_{R} such that

$$g_{\mathrm{L}} * h_{\mathrm{L}} * s = g_{\mathrm{R}} * h_{\mathrm{R}} * s \quad (2)$$

(we will omit the time index n if not necessary). Once g_{L} and g_{R} satisfy (2) for any speech signal s, the output of the CF is

$$z = g_{\mathrm{L}} * x_{\mathrm{L}} - g_{\mathrm{R}} * x_{\mathrm{R}} = g_{\mathrm{L}} * h_{\mathrm{L}} * s + g_{\mathrm{L}} * y_{\mathrm{L}}$$
$$- g_{\mathrm{R}} * h_{\mathrm{R}} * s - g_{\mathrm{R}} * y_{\mathrm{R}} = g_{\mathrm{L}} * y_{\mathrm{L}} - g_{\mathrm{R}} * y_{\mathrm{R}}. \quad (3)$$

The output of the ideal CF does not contain the contribution of s and provides information about the noise. The only problem is that the spectrum of the output z depends on g_{L} and g_{R} and can be seriously changed.

We introduce a vector-matrix notation where $\mathbf{X}_i$, $i \in \{\mathrm{L}, \mathrm{R}\}$, denotes the $L \times (N + L - 1)$ Toeplitz matrix whose first row and first column are $[x_i(1), \ldots, x_i(N), 0, \ldots, 0]$ and $[x_i(1), \ldots, 0]^T$, respectively. L is the length of filters g_{L} and g_{R} whose coefficients are stacked in vectors $\mathbf{g}_{\mathrm{L}}$ and $\mathbf{g}_{\mathrm{R}}$, respectively. Analogously, we define Toeplitz matrices $\mathbf{Y}_i$, $i \in \{\mathrm{L}, \mathrm{R}\}$, for signals y_i.

Assume now that x_{L} and x_{R} are noise-free recordings of the target signal. Common constructions of the CF [2] consist in fixing $\hat{\mathbf{g}}_{\mathrm{R}} = \mathbf{e}_D$ where $\mathbf{e}_D$ denotes the Dth column of the $L \times L$ identity matrix, D is an integer that determines the overall delay of the resulting CF, and finding $\hat{\mathbf{g}}_{\mathrm{L}}$ as

$$\mathrm{LS1:} \qquad \hat{\mathbf{g}}_{\mathrm{L}} = \arg\min_{\mathbf{g}_{\mathrm{L}}} \|\mathbf{g}_{\mathrm{L}}^T \mathbf{X}_{\mathrm{L}} - \hat{\mathbf{g}}_{\mathrm{R}}^T \mathbf{X}_{\mathrm{R}}\|_2^2. \quad (4)$$

A drawback of the above method, which is closely related to the transfer function ratio estimation in the frequency domain [15, 21], is that it does not take the impact of the resulting CF on the spectrum of the filter output into account.

In this paper, we propose a novel design of the CF which assumes that a target-free recording of a typical noise for the given environment is available. For now, let the recording be denoted by y_{L} and y_{R}, and x_{L} and x_{R} are the noise-free recordings of the target signal again. We propose to compute the CF according to

$$\mathrm{LS2:} \qquad \hat{\mathbf{g}}_{\mathrm{L}}, \hat{\mathbf{g}}_{\mathrm{R}} = \arg\min \|\mathbf{g}_{\mathrm{L}}^T \mathbf{X}_{\mathrm{L}} - \mathbf{g}_{\mathrm{R}}^T \mathbf{X}_{\mathrm{R}}\|_2^2$$
$$+ \epsilon \|\mathbf{g}_{\mathrm{L}}^T \mathbf{Y}_{\mathrm{L}} - \mathbf{g}_{\mathrm{R}}^T \mathbf{Y}_{\mathrm{R}} - \mathbf{y}\|_2^2 \quad (5)$$

where ϵ is a positive regularization parameter and $\mathbf{y}$ is the vectorized noise signal that we want to observe on the output of the CF. For example, $\mathbf{y}$ can be the vectorized signal $y_{\mathrm{L}}(n - D)$ where D is the delay parameter as in (4).

Similarly to (4), the criterion in (5) is quadratic also. The minimizer is given by

$$\begin{bmatrix} \hat{\mathbf{g}}_{\mathrm{L}} \\ \hat{\mathbf{g}}_{\mathrm{R}} \end{bmatrix} = \mathbf{W}^{-1}\mathbf{h} \quad (6)$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{X}_{\mathrm{L}} \\ -\mathbf{X}_{\mathrm{R}} \end{bmatrix} [\mathbf{X}_{\mathrm{L}}^T, -\mathbf{X}_{\mathrm{R}}^T] + \epsilon \begin{bmatrix} \mathbf{Y}_{\mathrm{L}} \\ -\mathbf{Y}_{\mathrm{R}} \end{bmatrix} [\mathbf{Y}_{\mathrm{L}}^T, -\mathbf{Y}_{\mathrm{R}}^T]$$

$$\mathbf{h} = \epsilon \begin{bmatrix} \mathbf{Y}_{\mathrm{L}} \\ -\mathbf{Y}_{\mathrm{R}} \end{bmatrix} \mathbf{y}. \quad (7)$$

Note that $\mathbf{W}$ is a symmetric block-Toeplitz matrix with blocks of size $L \times L$. An efficient solver of (6) is the block Levinson-Durbin algorithm derived in [23] whose complexity is $\mathcal{O}(dL^2)$ where d is the number of blocks (here $d = 2$).

The scale of the solution (6) depends on ϵ and on the norm of $\mathbf{y}$. It is therefore handy to normalize the solution so that the
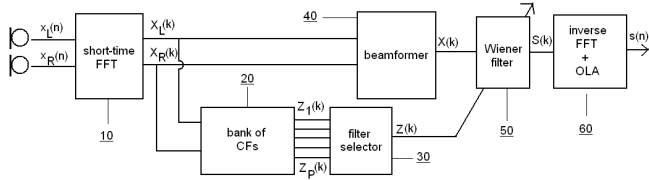
Fig. 2. Scheme of the Noise Reduction System

output of the resulting CF applied to the pure noise yields a variance equal to the input variance.

After having the cancellation filters prepared in time domain, they can be transformed to the frequency domain and stored for a further usage in the memory of the mobile phone.

3. NOISE REDUCTION SYSTEM

The proposed noise reduction scheme is drawn in Fig. 2. Each block of input signals is processed in parallel by all cancellation filters in the bank 20. The next step is a filter selector 30, which selects the filter whose output yields minimum variance. In general, this output need not have the least speech leakage. Nevertheless, the selection is reasonable since the portion of energy corresponding to the speech is usually large (the speaker is close to microphones). A more sophisticated but complex approach was proposed, e.g., in [26]. Outcome of the selected filter is taken as an estimate of the noise signal.

The upper branch of the scheme contains a beamformer 40, which provides an initial estimate of the speaker voice. In the case when one microphone is located on the front side of the phone and the second one is on the rear side, the signal from the former microphone is taken as output of the beamformer. In case that both microphones are on the front side, the one yielding higher variance (because it could be closer to the speaker) can be used.

The next step consists in subtraction of the estimated noise signal from the initial estimate of the target in 50. Here we use a simple spectral subtraction method based on the frequency-domain Wiener filter with the noise gain parameter τ [27], but a more sophisticated methods could be used such as the double spectral subtraction [28] or PLD from [4, 5]. In order to improve the perceptual quality of the final output, a frequency-domain smoothing [29] can be employed for frequencies higher than certain threshold.

4. EXPERIMENTS

For our experiments, we have developed a model of a dual-channel mobile phone. It consists of a printed circuit board with three integrated microphones that are used, e.g., in Sony Ericsson K850. Two microphones are placed in the front bottom corners and one is placed in the top left corner on the rear side (see Fig. 3). The left-hand side microphones are
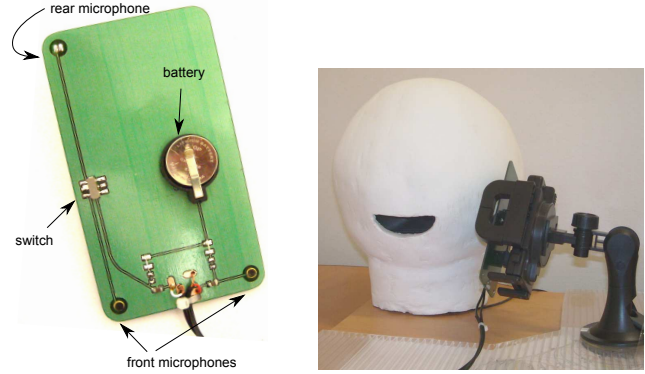

Fig. 3. Model of a mobile phone and an artificial head used in experiments.

switchable, and we test the two corresponding dual-channel arrangements. Signals from the selected microphones are amplified by M-Audio AudioBuddy pre-amplifier and recorded by M-Audio Profire 2626 external sound card. The sampling frequency is 16 kHz.

Our development and testing scenario consists of an artificial head made of gypsum (see Fig. 3). A loudspeaker is placed inside the head and directed towards a hole to simulate mouth. All experiments were done in a room having the reverberation time about $T_{60} = 300$ ms. Speakers are simulated using signals taken from the TIMIT database. Stereo signals of a diffuse babble and traffic noise were taken from [24].

We derived several banks of CFs for the artificial head. Each bank contained 14 CFs for different positions of the mobile around the artificial mouth. The mobile was mounted in a stand as shown in Fig. 3. Training noise-free recordings each of length 4 s were obtained by playing training utterances from the artificial head.

Two different speakers (male and female) and two microphone arrangements were considered (two front microphones or one front and one rear microphone), and two approaches LS1 and LS2 were used to compute the CFs of length 1000 with the delay parameter $D = 20$. In total, eight banks were derived. The variants of the proposed method using the corresponding banks will be denoted LS1 and LS2, respectively.

Testing target signals were recorded from the artificial head placed in a different location in the office room than for the training. They contain utterances of length 7.5 s of the same speakers as for the training. During the recordings, the model of the mobile was moved around the mount of the artificial head. The mobile was not mounted in the stand as for the training but was held in hand of the first author.

As noise signals, we used babble and traffic noise but also an uttering man, whose speech was played by a loudspeaker that was placed one and half meter in front of the artificial head. The noise signals were mixed with the testing signals at a ratio between -10 and 10 dB (*input SNR*).

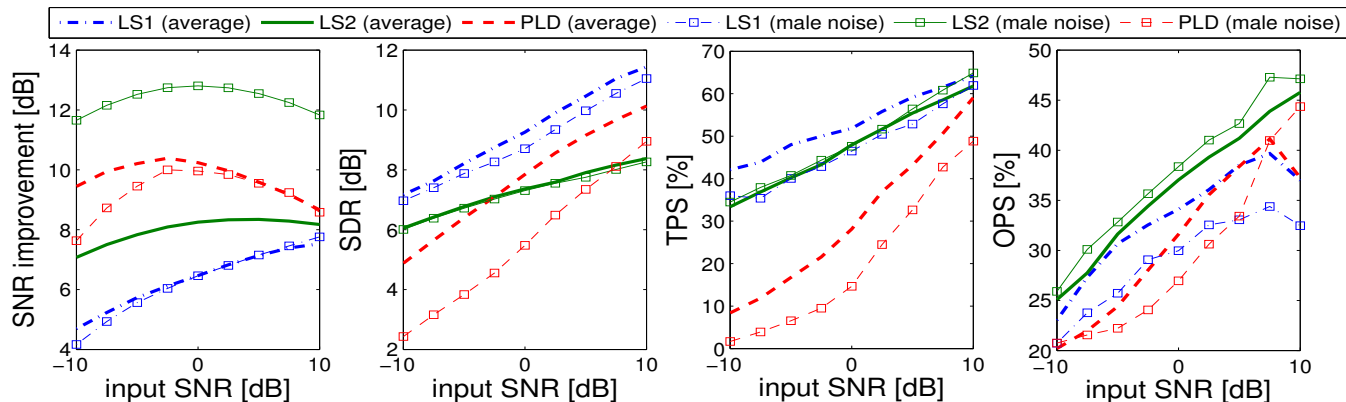To measure the performance of the target cancellation

Fig. 4. Results achieved by two variants of the proposed method (LS1 and LS2) and by the PLD algorithm [4] when using the front and rear microphones.
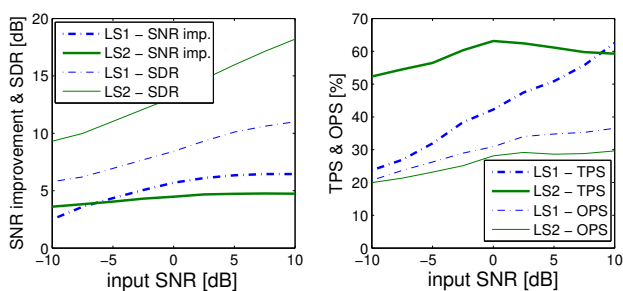

Fig. 5. SNR improvement, SDR, TPS and OPS for the setup with two front microphones.

within the proposed method, we evaluate the Noise-to-Signal Ratio (NSR) which is the ratio of energy of the target and noise contributions at the output of the blocking matrix.

The enhanced signals at the output of the noise reduction methods are evaluated in terms of Signal-to-Noise Ratio (SNR) and Signal-to-Distortion Ratio (SDR). SNR measures the residual noise in the enhanced signal while SDR reflects the damage of the target signal in it. Perceptual quality is evaluated in terms of Target-related Perceptual Score (TPS) and Overall Perceptual Score (OPS) computed using the PEASS software version 2.0 [25].

We conducted experiments with many options; detailed results are available on a web site[1]. Here, we present the results for the case when the bank of CFs was derived from training signals of the male speaker while the testing speaker was female. First, we consider the setup with the front and rear microphone. The LS2 variant is tuned for the noise of the male speaker; $\epsilon = 0.5$ in (5).

The same arrangement of microphones is assumed by the Jeub's PLD algorithm [4], so we compare it using the same parameters as in [4]. Results averaged for the babble, traffic and male speaker noise and separately for the male noise are shown in Fig. 4.

---

[1] http://itakura.ite.tul.cz/zbynek/downloads.htm

In this example, PLD achieves higher SNR but significantly lower SDR, TPS and OPS compared to LS1 and LS2. The distortion of the target signal is mainly caused by the leakage of the target signal to the noise reference signal (or to its estimated power spectrum). PLD relies on a sufficient attenuation of the speaker voice on the rear microphone, while the proposed methods improve the voice attenuation by the bank of CFs, which is more efficient. In case of the babble noise, LS1 and LS2 improve the NSR at the blocking matrix output on average by 9.4 dB and 8.8 dB, respectively, while the NSR on the rear microphone is only by 5.7 dB better than on the front microphone. This phenomenon is significant mainly when input SNR is lower than 0 dB.

Note that the performance of LS2 is superior in case of the male speaker noise. It demonstrates the effect of the adjustment of the bank of CFs to the noise.

In the second example, we tested the setup with two front microphones, which is not suitable for PLD. Therefore we compared LS1 and LS2 only. The bank of CFs in LS2 was tuned for the babble noise. Results in Fig. 5 show that LS2 achieves better SDR and TPS than LS1 due to the adaptation to the babble noise. On the other hand, SNR by LS1 is slightly higher than that by LS2, which finally leads to the better OPS.

Comparing the results in Figures 4 and 5 indicates that the system with one front and one rear microphones reduces the noise better, especially, in terms of the SNR improvement and OPS. The rear microphone provides a better starting point to obtain a good noise reference signal. On the other hand, two front microphones may be more attractive option for other applications such as stereo recording.

5. CONCLUSIONS

We have proposed a new method for noise reduction in dual-microphone mobile phones and a novel construction of target cancellation filters. The arrangements of the microphones can be arbitrary. In comparison to PLD, it achieves better perceptual quality and is able to work in difficult scenarios where SNR is lower than 0 dB.

6. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, 1st edition, Springer-Verlag, Heidelberg, 2005.

[2] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model", *Proc. of WASPAA 2009*, pp. 133 - 136, New Paltz, New York, Oct. 2009.

[3] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 419–422, 1997.

[4] M. Jeub, C. Herglotz, C. M. Nelke, C. Beaugeant and P. Vary, "Noise Reduction for Dual-Microphone Mobile Phones Exploiting Power Level Differences, ", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1693–1696, Kyoto, Japan, Mar. 2012.

[5] N. Yousefian, A. Akbari and M. Rahmani, "Using power level difference for near field dual-microphone speech enhancement," *Applied Acoustics*, vol. 70, pp. 1412–1421, 2009.

[6] J. Hu and M. Lee, "Speech Enhancement for Mobile Phones Based on the Imparity of Two-Microphone Signals, " *Proceedings of the 2009 IEEE International Conference on Information and Automation*, pp. 606–611, Zhuhai/Macau, China, 2009.

[7] K. Li, Y. Guo, Q. Fu, J. Li, and Y. Yan, "Two Microphone Noise Reduction Using Spatial Information-Based Spectral Amplitude Estimator," *IEICE Trans. Information and Systems*, vol. E95-D, no. 5, pp. 1454–1464, May 2012.

[8] N. Yousefian and P. C. Loizou, "A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 2, Feb. 2012.

[9] Z. Zhang and M. Etoh, "ICA-based Noise Reduction for Mobile Phone Speech Communication," *Proceedings of 16th International Conference on Computer Communications and Networks*, pp. 470–473, Aug. 2007.

[10] H. Sawada, S. Araki, R. Mukai, S. Makino, "Blind Extraction of Dominant Target Sources Using ICA and Time-Frequency Masking," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165–2173, Nov. 2006.

[11] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[12] O. Hoshuyama, A. Sugiyama, A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol.47, no. 10, pp. 2677–2684, Oct. 1999.

[13] W. Herbordt, W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broad-band signals," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, vol. 4, pp. IV-4187, May 2002.

[14] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, pp. 4–24, Apr. 1988.

[15] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[16] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, Jan. 2011.

[17] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[18] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Information Theory*, vol. 40, no. 2, pp. 340-349, 1994.

[19] N. Levinson, "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.

[20] Y. Lin, J. Chen, Y. Kim and D. Lee, "Blind channel identification for speech dereverberation using ℓ_1 norm sparse learning," *Advances in Neural Information Processing Systems 20*, pp. 921–928, MIT Press, 2008.

[21] O. Shalvi and E. Weinstein, "System identification using non-stationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055-2063, Aug. 1996.

[22] I. Tashev, *Sound Capture and Processing: Practical Approaches*, John Wiley & Sons Ltd., 2009.

[23] H. Akaike, "Block Toeplitz Matrix Inversion," *SIAM Journal on Applied Mathematics*, vol. 24, no. 2, pp. 234-241, March 1973.

[24] ETSI 202 396-1, *Speech and multimedia Transmission Quality (STQ); Part 1: Background noise simulation technique and background noise database*, 03 2009, V 1.2.3.

[25] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, "Subjective and objective quality assessment of audio source separation, " *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sept. 2011.

[26] J. Málek, Z. Koldovský and P. Tichavský, "Semi-Blind Source Separation Based on ICA and Overlapped Speech Detection", *Proc. of The 10th International Conference on Latent Variable Analysis and Source Separation (LVA/ICA 2012)*, LNCS 7191, pp. 462-469, Tel-Aviv, Israel, March 12-15, 2012.

[27] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Tr. Acoust. Speech and Signal Proc.*, vol. 27, pp. 113–120, 1979.

[28] H. Gustaffson, I. Claesson, S. Nordholm, and U. Lindgren, "Dual microphone spectral subtraction," *Tech. Rep., Department of Telecommunications and Signal Processing, University of Karlskrona/Ronneby*, Sweden, 2000.

[29] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.